Réunion 5 Novembre 2019

- 18% des données ont un magasin,
  - pas suffisant pour faire une étude approfondie
  - certains magasins sont sous représentés -> biais

Liste de pourcentage d'entrées non NaN par catégorie ci-joint. Certaines entrées sont non NaN mais "0.0", "Unknown", "To be completed", et comptent pour des NaNs

On remarque que l'on retenir beaucoup d'aliments en fonction de leur catégorie ("categories","pnns_groups","main_category") et aussi leurs apports nutritionnels (fat, salt, energy, carbohydrates, sugar, proteins,…) ainsi que additifs et traces allergiques.

=> Une étude portée sur **la comparaison nutritive de régimes alimentaires** (sans-gluten, vegan, bio,…) serait plus complète et facile à construire.

- possible extension à une comparaison **autre** que nutritive ?

<u>Pistes à explorer:</u>

- Créer une ou plusieurs catégories sur le dataset pour catégoriser les aliments en fonction de leur affiliation à leur régime et/ou catégorie. Voir combien de données sont traitables et avec quelle fidélité.

- Voir si il y une correlation entre l'absence d'information dans une entrée et le timestamp de l'insertion de la donnée. Si oui, on peut dire que nos études portent sur les dernières X années, si elles comportent la grande majorité des données (pour ne pas en perdre).

- Voir une extension possible de la problématique avec un autre dataset.